NWC TP 7027

Multisensor Evaluation Framework

by
David C. Foyle
Aircraft Weapons Integration Department

SEPTEMBER 1989

NAVAL WEAPONS CENTER
CHINA LAKE, CA 93555-6001

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
September 1989

3. REPORT TYPE AND DATES COVERED
Summary; May 1986 - August 1989

4. TITLE AND SUBTITLE
Multisensor Evaluation Framework

5. FUNDING NUMBERS

6. AUTHOR(S)
David C. Foyle

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Weapons Center
China Lake, CA 93555-6001

8. PERFORMING ORGANIZATION REPORT NUMBER
NWC TP 7027

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Human Factors Block Funding, NPRDC and NADC

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

(U) Based on existing integration models in the literature, an evaluation framework is developed to assess the operator's ability to use multisensor, or sensor fusion, displays. In general, the multisensor display can be evaluated by comparing the operator's performance with the multisensor display to predicted performance derived from the use of the individual sensors comprising the display. An experiment demonstrating the usefulness of the proposed evaluation framework was conducted.

14. SUBJECT TERMS
Display evaluation
Information integration
Information transfer
Multisensor display
Sensor fusion display

15. NUMBER OF PAGES
24

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UL

CONTENTS

ACKNOWLEDGMENT

The author wishes to thank the following people for their help with analysis and programming: David Gemert, Howard MacCauley, Susan Savage, and Steven Wilson. Data collection and preliminary analysis were conducted at the Naval Weapons Center and supported by the Office of Naval Technology 6.2 Human Factors Block. Remaining analyses and writing of this manuscript were supported by the National Aeronautics and Space Administration (NASA) Ames Research Center, and conducted after the author moved to NASA Ames Research Center. Thanks to Michael Brickner, Marion Kibbe, and Michael Shafto for reviewing an earlier version of this paper. The viewpoints in this paper represent those of the author and not necessarily those of any government agency.

INTRODUCTION

Many different types of imaging sensors exist, each sensitive to a different region of the electromagnetic spectrum. Passive sensors, which collect energy emitted or reflected from a source, include television (visible light), night vision devices (intensified visible light), and infrared (heat) sensors. Active sensors, in which objects are irradiated and the reflected energy from those objects collected, include sonar and ultrasound (acoustic waves) and radar (radio waves).

These sensors were developed because of their ability to increase the probability of identification or detection of objects under difficult environmental conditions. Each sensor is sensitive to different portions of the spectrum; therefore resultant images contain different information even when used under the same conditions. Because of this variety, image processing algorithms that will "fuse" the information from more than one sensor into a single coherent display image are being developed. These displays are termed multisensor or sensor fusion displays.

The work described in this paper was conducted to guide the development of such multisensor displays. An engineer developing such a system constantly reviews the resulting display on a subjective basis. More formal testing is also necessary. Suppose, for example, that two sensor sources are available to operators and that each of these sensors alone leads to 0.70 probability of target recognition under some particular environmental conditions. What is the expected probability of target recognition when the two sensors are combined according to some image processing technique? If observed target recognition improves to 0.80 with a sensor fusion system, is that a large improvement, or should one actually expect much more? The ability to answer these types of questions can lead to a better human-machine system that can be evaluated both relatively and absolutely: relatively, by determining which systems are better than others, and absolutely, by comparing operator performance to theoretical expectations.

INFORMATION INTEGRATION MODELS

Previous work has been conducted on the topic of how operators integrate the information from multicomponent auditory signals (Reference 1), from the visual and auditory senses (Reference 2), and from multiple observations over time (Reference 3). These models all predict operator integration performance as a function of the operator's performance with the individual stimuli comprising the integration task. Two classes of models have been developed: decision combination models and observation integration models (for a review, see Reference 4). The decision combination models assume that in the integration task the operator makes an individual decision about each aspect of the combined display and then combines those decisions to yield one final decision. At the time of the final decision, only the previous decisions are available and not the information that led to the individual decisions. The observation integration models, on the other hand, assume that the operator does have access to that information. The internal representations of the individual observations (e.g., likelihood ratios) are then combined, yielding only one decision.

The simplest version of a decision combination model is the probability summation, or statistical summation, model. As Reference 4 notes, it is derived from the independence theorem of probability theory and was first proposed by Pirenne as a perceptual model (Reference 5). It states that performance with a complex stimulus is predictable from the performance with the individual stimuli according to the following equation:

$$p_{12} = p_1 + p_2 - p_1 p_2$$

where p_1 and p_2 represent detection probabilities for the two stimuli presented in isolation, and p_{12} is the detection probability when both stimuli are available.

The most cited version of the observation integration model is derived from the theory of signal detectability and was originally proposed by Green (Reference 1). As in Pirenne's model (Reference 5), in its most simple form, the information from the two sources is also assumed to be independent and uncorrelated. The model is stated in terms of the sensitivity measure, d':

$$d'_{12} = \left[(d'_1)^2 + (d'_2)^2 \right]^{1/2}$$

where d'_1 and d'_2 and d'_{12} respectively, represent performance with the two stimuli presented in isolation, and when both stimuli are available.

Swets has noted that the statistical summation model fits simple detection data fairly well when the observed detection probabilities are corrected for chance success (Reference 4). Similarly, in the experiments in which it has been applied, the observation integration model well represents the data. In general, the statistical summation model predicts better integration performance than the observation integration model presented here. My calculations indicate that when both models are expressed in corrected-for-chance probability of a correct response, the statistical summation model predicts the detection probability for integration to be about 0.05 (depending on the absolute level) higher than that predicted by the observation integration model.

Various versions and extensions of these models have been proposed. These include various rules of decision combination (Reference 3), correlation (informational redundancy) among inputs for decision combination (Reference 2), and observation integration (Reference 6), as well as versions of the observation integration model in which the separate inputs are differentially weighted (References 7 and 8).

The two integration models presented here have been incorporated into the development of a framework to evaluate combined human-machine performance for sensor fusion displays. Additionally, the framework could be used to evaluate an operator's ability to integrate information from a multiple monitor display system or a screen paging system.

A PROPOSED EVALUATION FRAMEWORK

A sensor fusion display typically refers to the combined image display resulting from the application of one image processing technique on two or more individual sensor images. The proposed framework for evaluating the operator's ability to use such systems is considered a normative approach; the operator's performance with the sensor fusion display can be compared to performance on the individual sensor displays comprising that display and to various optimal models of integration.

Typically, as the environmental conditions change in which the individual sensor operates, so does the information content of that image. The information content of the image can be "scaled" by the operator's ability to perform a target identification or discrimination task. One would expect task performance with a sensor fusion display formed from two low information content (hence, poor-performance) images to still be relatively poor. Similarly, two high information content (high-performance) sensor images should yield good performance when combined into a sensor fusion display. Assuming that there was some independent information in the two individual sensor images, one would also expect performance with the sensor fusion display to be better than with either of the two individual sensors alone. This results in a three-dimensional performance space: performance with the sensor fusion image is a function of the performance levels associated with the two individual sensor images.

Figure 1 shows part of this performance space associated with a sensor fusion display. The abscissa and the ordinate result from the stimulus-performance scaling for sensor (or display) 1 and sensor (or display) 2, respectively, when viewed by an operator in isolation. The figure shows the iso-performance horizontal "slice" through the space in which all performance data points represent 0.72 (corrected for chance) target recognition probability when the two sensor sources are combined into a sensor fusion display and presented to an operator. As noted, the actual performance space is three-dimensional and is represented in Figure 2 by similar-appearing "slices" at three performance levels. Data points A, B, C, and D are discussed below.

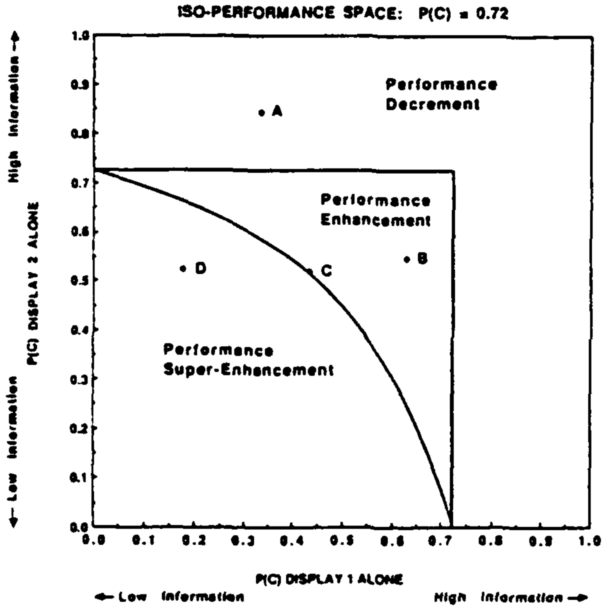


FIGURE 1. A Proposed Evaluation Framework for Multisensor Displays.

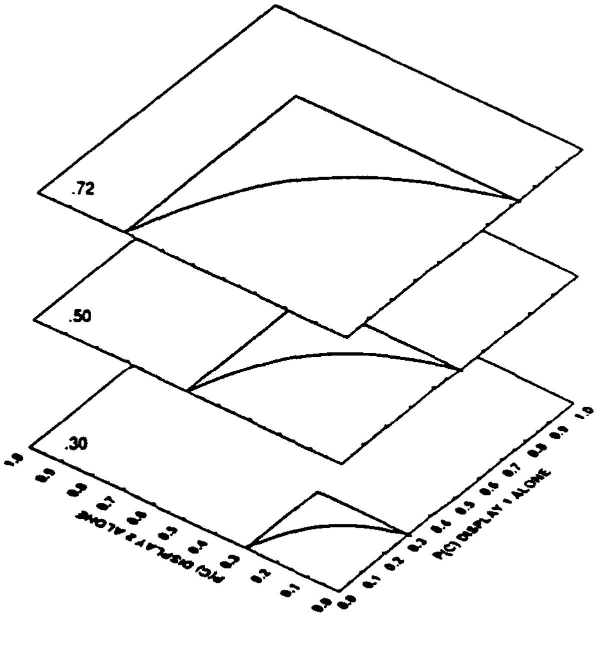


FIGURE 2. Three Horizontal Slices Through the Three-Dimensional Performance Space. The number on each overlay represents the performance level, in P(C), for the dual display or multisensor display task.

Because the sensor fusion display data are plotted as iso-performance slices, data points near the origin represent better performance than away from the origin. For the same level of performance, a data point near the origin represents a condition in which very little information was available in the two displays, whereas a data point away from the origin represents a condition in which relatively more information was available in the separate displays. Thus, data points near the origin represent increased operator integration efficiency. In these figures and all remaining references, P(C) refers to the proportion of correct responses with a correction for chance applied. A correction for chance is necessary when measuring performance in P(C) units because the integration models require that a performance level of zero be associated with the operator receiving no information from the display. No such correction is necessary when measuring performance in d' units since d' = 0 refers to chance performance.

As can be seen from the two figures, the sensor fusion performance space can be divided into three separate areas, Performance Decrement, Performance Enhancement, and Performance Super-Enhancement, each with unique interpretations if data points lie in those areas. The two right-angle lines dividing the Performance Decrement and Performance Enhancement areas are determined by the horizontal and vertical lines crossing the axes at the level of performance (P(C) = 0.72 in Figure 1) for the sensor fusion display. The smooth curves separating the Performance Enhancement and Performance Super-Enhancement areas are the predictions of the statistical summation model (see above) where $p_{12} = 0.72$ in Figure 1 and 0.30, 0.50, and 0.72 in Figure 2. (For clarity, only the statistical summation model curve has been shown in the two figures. Because the research to date does not favor either the statistical summation or observation integration models, both predictions will be used when evaluating the experimental data.)

The interpretation of the data points falling into the three areas is best illustrated by example.

PERFORMANCE DECREMENT

Suppose under a given environmental condition, an operator achieved target recognition performance of P(C) = 0.33 when viewing Sensor 1 in isolation and P(C) = 0.84 when viewing Sensor 2 in isolation. When these two sources are both available (separately on two monitors, or fused on a single monitor according to a sensor fusion algorithm) to the operator and performance is P(C) = 0.72, the resultant data point would be the one labeled "A" in Figure 1. Obviously, in this situation, the operator has not improved his overall targeting performance. In fact, performance in the combined display case has now decreased to P(C) = 0.72, whereas previously the operator used Sensor 1 in isolation and reached a performance of P(C) = 0.84. Such a performance decrement could be the result of the deletion of necessary information by the sensor fusion algorithm or could represent a cognitive limitation on the part of the operator.

PERFORMANCE ENHANCEMENT

Data point "B" in Figure 1 would result if P(C) = 0.72 performance obtained in the combined case, when Sensors 1 and 2 yielded P(C) = 0.63 and P(C) = 0.55, respectively, in isolation. In this case, performance improved since the operator did better in the combined case (0.72) than with either of the two sources alone (0.55, 0.63). However, one model of information integration, the statistical or probability summation model, predicts a larger improvement in this case. Thus, for data points falling in this region, there is some performance improvement, but one would expect more. In fact, data point "C," lying on the statistical summation model curve, shows that the model predicts that if Sensor 2 performance was 0.52, Sensor 1 performance need only be 0.42 to result in combined performance of 0.72.

Operator performance occurring in this region would occur when some of the information in the two sources is redundant (correlated and not independent), or when the operator or the sensor fusion algorithm integrate the information, but do so suboptimally. The statistical summation model (as well as the observation integration model) can be viewed as an upper limit of integration: it assumes that the information in the two sources is independent and non-redundant, and does not assume any decrease in performance due to the limits of cognitive processes (i.e., memory limitations, work load, or suboptimal decision strategies).

PERFORMANCE SUPER-ENHANCEMENT

Data point "D" in Figure 1 would result if a combined performance of P(C) = 0.72, and individual performance for the two sensors was P(C) = 0.17 and P(C) = 0.52. Data points falling in this region between the model prediction and the origin represent improved performance that is better than is predictable from the model. That is, when the two sources of information are viewed by the operator, some new, previously unusable, information emerges that results in much better performance.

The random-dot stereogram display (Reference 9) can be thought of as an example of a sensor fusion display that has these properties. In these displays, random dots are offset differentially, yielding a perception of an object in the third dimension. In such a stereogram there is no information whatsoever in the individual halves of the stereogram. The information is represented as differences between the two displays. The object is observable only by stereoscopically fusing the two halves of the stereogram or analytically determining the differences. In fact, if one conducted an experiment in which subjects had to state the "floating" shape, one would presumably obtain chance performance when viewing only one stereogram half and perfect performance when both stereogram pairs are viewed. This represents Performance Super-Enhancement because based on chance performance with the stereogram halves, one would conclude that they contain no information. This would lead one to predict chance performance when both halves are available, which obviously is not the case. Clearly, conditions in which Performance Super-Enhancement occur could be capitalized upon to produce useful sensor fusion techniques. The proposed evaluation framework provides for the ability to recognize and quantify such conditions.

USE OF THE EVALUATION FRAMEWORK

To evaluate human performance with a proposed sensor fusion system using the proposed evaluation framework, the following steps must be taken.

1. **Performance Scaling of Sensor 1.** Determine the psychometric function relating task performance (target/non-target or m-alternative forced choice) to the environmental conditions of interest. For example, infrared imagery is degraded by increasing atmospheric moisture. The information content of each sensor image varies with the environmental conditions, and in a sense, this *scaling* estimates the amount of information available to the operator with Sensor 1 alone under those conditions.

2. **Performance Scaling of Sensor 2.** Similar to Sensor 1.

3. **Performance with Sensor Fusion Display.** For various combinations of environmental or sensor conditions previously evaluated in isolation, determine task performance using the proposed fusion algorithm and associated display.

4. **Performance with Operator Integration.** As in the sensor fusion evaluation phase, determine task performance with both sensors but with either two displays or a split screen. This conditicn acts as a control condition, essentially allowing the operator to integrate the information from the two sensors. A sensor fusion algorithm should yield better task performance than when the operator uses two displays or a split-screen display.

An experiment was conducted to demonstrate the usefulness of the proposed "evaluation framework." Since this work was aimed at the general methodology of evaluating sensor fusion displays and due to the unavailability of a sensor fusion system that operates on imaging sensors and yields a "fused" image, step no. 3

above (performance with a sensor fusion algorithm) was not performed. In this particular experiment, two displays with independent samples of the stimuli simulating one type of sensor were used. (One might view this as a simulation in which sensor fusion operates on image samples collected at different times or from two data-linked sensor platforms using the same sensor type.) This is a special case of the sensor fusion condition, but the application of the evaluation framework to two different sensors is identical. Because there was only one type of sensor to scale, step no. 2 above (performance scaling of Sensor 2) was not necessary. That step, of course, would have been necessary if two different sensors were used.

METHOD

SUBJECTS

Four volunteer subjects were tested. The author (subject no. 1 reported herein) and three colleagues were tested for 8 to 12 hours each. All subjects reported normal or near-normal close-distance corrected vision.

SHIP IMAGES

Side profiles of ships from *Jane's Fighting Ships* (Reference 10) were used as stimuli in the study. The ship images were approximately 2.2 centimeters high by 6.6 centimeters wide and presented on Setchell Carlson 10M915 CRT displays driven by a Genisco GCT-3000 image system. Viewing distance was determined by the subject. The ship images were digitized profiles composed of 60 points connected by lines, evenly-spaced in the horizontal dimension, as shown in Figure 3. Each undistorted image spanned 60 vertical pixels by 180 horizontal pixels on the CRT display.

The independent variable was the amount of noise added to the vertical dimension of the ship profiles. This variable was quantified as "sigma," the standard deviation of a Gaussian distribution with mean zero. A computer algorithm based on random number samples was implemented for this purpose. For a given sigma level, 60 numbers (both positive and negative) were drawn from that distribution and added individually to the vertical pixel value of each of the 60 points of the ship profile. For example, for sigma = 5, on average 68% (the area of the Gaussian curve from -1 to +1 standard deviation) of the points would be within 5 pixels of the original undistorted vertical value. New numbers were drawn from the distribution for each ship and each trial; thus no two distortions were identical. Two of the ships with distortions for four levels of sigma are shown in Figure 4.

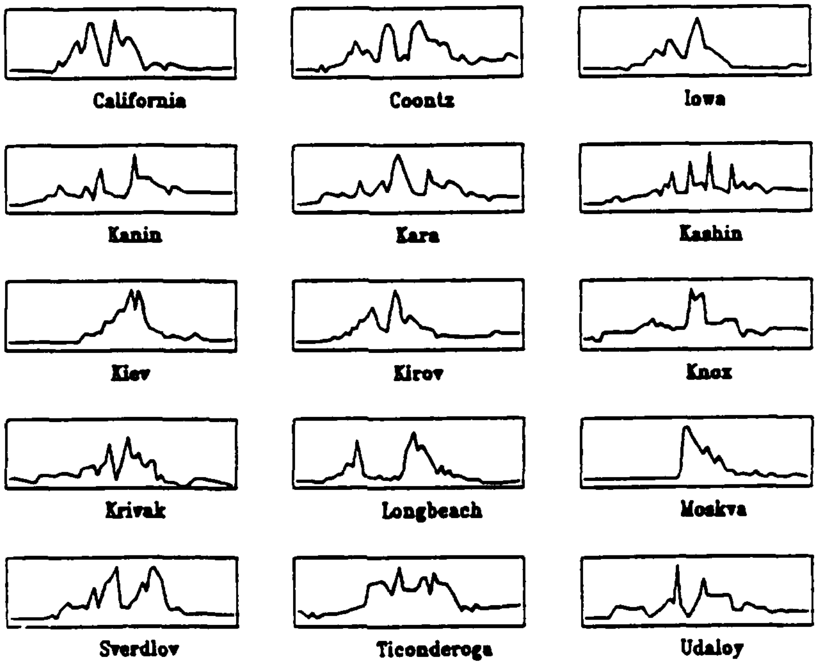
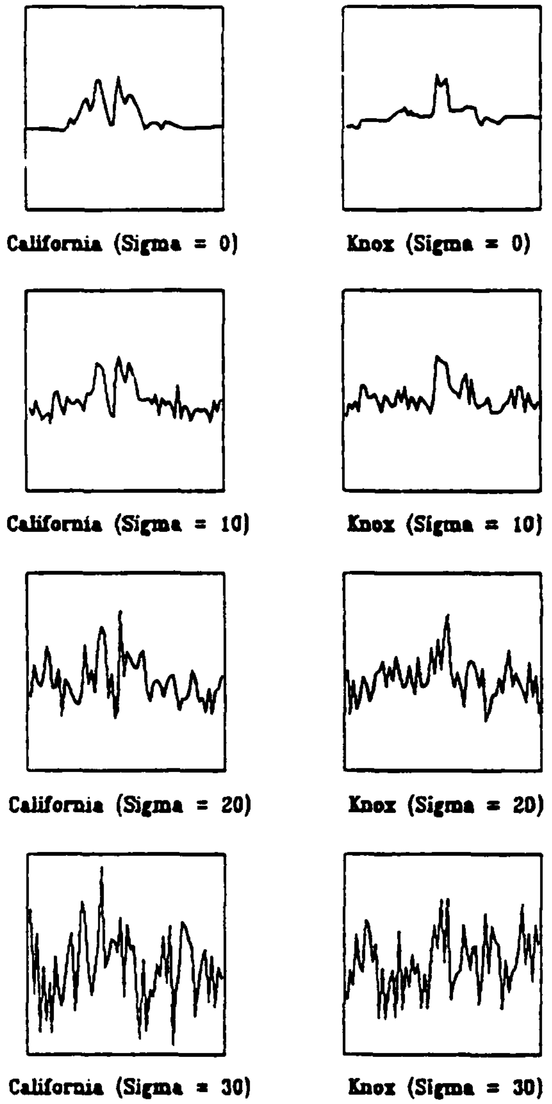


FIGURE 3. Digitized Ship Profiles of the 15 Ships Used in the Experiment.

SINGLE-DISPLAY TASK

The task was a 4-alternative, forced-choice (4AFC) task and was controlled by a VAX 11/750. Subjects viewed a single-CRT display partitioned by vertical and horizontal lines into four numbered quadrants, each containing a ship image. Subjects were required to identify which of the four ships was the *California*, which was always present in each display. In addition to the target ship, *California*, three different distractor ships were presented that were randomly drawn from the 14 remaining ships shown in Figure 3. During testing, subjects were allowed to study the ship images for an unlimited time. Subjects responded with a button press of the numbers 1 to 4, referring to the quadrant which they believed held the target ship, *California*. Full feedback (quadrant responded, correct/incorrect, and the correct quadrant if incorrect) was given after each trial via a digitized speech capability on a Texas Instrument (TI) Portable Professional computer. The end of the verbal feedback initiated the next trial. All four ships on a display were created with the same sigma (distortion) level.



California (Sigma = 0)

Knox (Sigma = 0)

California (Sigma = 10)

Knox (Sigma = 10)

California (Sigma = 20)

Knox (Sigma = 20)

California (Sigma = 30)

Knox (Sigma = 30)

FIGURE 4. Examples of the Distortion Algorithm Used for the Ship Profiles.

DUAL-DISPLAY TASK

Conditions were identical to the single-display task with the following exceptions. Ship images were presented on two horizontally adjacent CRT displays. The four ships were arranged in the same spatial pattern on both displays (i.e., if the *California* was in quadrant 2, it was at that location in both displays, and likewise for the 3 remaining distractor ships). All eight ships presented were independently distorted by new draws from the Gaussian distribution. Drawing the distortion values from independent Gaussian distributions ensures that the information on the two displays was independent and uncorrelated (see Reference 11 for a discussion of this technique). Within each display on a given trial, the four ships had a constant value of sigma. The value of sigma for each display was determined as described below.

TEST PROCEDURE

Trials were tested in 25-trial groups, or blocks. Full performance feedback (number correct as a function of sigma) was given after each block. Single- and dual-display tasks were alternated approximately every two blocks so that the first few blocks of each task could be discarded as practice.

For the single-display task, five levels of sigma were tested for each subject and psychometric functions were generated. The five sigma levels used in the first few blocks contained both low and moderate levels of sigma (from 0 to 20 or 5 to 25). Coupled with feedback after each trial, subjects were able to learn the salient features of the target ship relatively quickly.

For the dual-display task an adaptive threshold estimation procedure was used that mathematically converged on the stimulus value associated with corrected-for-chance performance of P(C) = 0.72 (Reference 12). In this technique, display 1 contained ships of one of five previously selected sigma levels. In display 2, the level of sigma (in increments of 5 units) of the ships varied as determined by the threshold estimation procedure. This procedure determined the display 2 sigma level as a function of the history of the subjects' responses. This resulted in five interleaved adaptive tracks (one associated with each level of sigma in display 1). This procedure results in five pairs of sigma levels for displays 1 and 2 that all yield P(C) = 0.72 target identification performance in the limit. When plotted as in Figures 1 and 2, the data points all lie on the same horizontal "slice" through the three-dimensional performance space (specifically the one labeled ".72" in Figure 2).

Another procedure to estimate a stimulus value associated with a constant level of performance is to collapse each of the interleaved tracks across trials into a psychometric function that can then be fit with a curve. This procedure has proven to be useful and is an efficient method to concentrate observations in the performance range of interest (Reference 13). This procedure was attempted in this experiment but did not yield reliable estimates because of the number of data points in the psychometric function. This procedure would be preferred to the analysis that was used if the number of data points allowed the estimation of stable psychometric functions.

The use of the tracking algorithm in the present study should be viewed as an experimental convenience, with its merits or limitations tangential :o the use and development of the evaluation framework. Use of the tracking algorithms may not be appropriate in the actual evaluation of a specific proposed sensor fusion system because only certain discrete pairs of stimulus combinations may be possible. For example, if atmospheric humidity affects the image quality of both sensors, only combinations of stimuli in which the atmospheric humidity was identical make sense. In those cases, one would not use the threshold estimation procedure but would test the imagery associated with the environmental conditions of interest. These resultant data pairs would then be plotted on the various appropriate horizontal slices similar to Figure 2 (determined by the dual-display or fusion-display performance). Interpretation of the placement of the data points on the various horizontal slices would be carried out in a fashion similar to that previously described.

RESULTS

In the single-display condition, the first four blocks of data collected were eliminated from the analysis as practice data. This decision was made after viewing the accuracy data for each block. The data were analyzed and summarized as both d' and P(C) (corrected for chance) measures. The d' data were converted to ln(sigma) by ln(d') and fit with a regression equation weighted by the number of observations per point (References 13 and 14) as shown in Figure 5. The P(C) values reported have been corrected for chance according to the following equation (see Equation 5.4 in Reference 3):

$$P(C)_{\text{CORRECTED}} = \left[P(C)_{\text{OBSERVED}} - 0.25 \right] / 0.75$$

These P(C) data were converted to ln(sigma) by z-score (unit-normal deviate) and likewise fit with a weighted regression equation. The data and associated psychometric functions are shown in Figure 6. Each data point in the two figures is based on 60 to 100 observations. The Pearson correlations associated with the linear regressions ranged from -0.91 to -0.99.

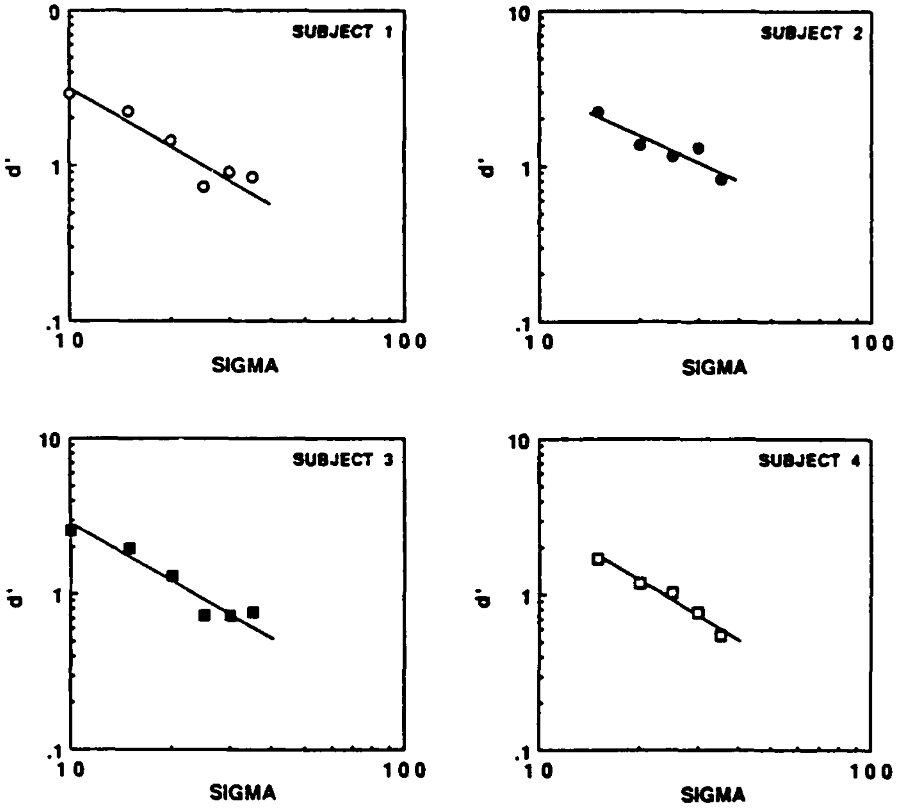


FIGURE 5. Psychometric Functions for the Four Subjects Relating Distortion Level (sigma) to the Sensitivity Measure, d', in the Single-Display Task.

In the dual-display condition, the five interleaved adaptive tracks were analyzed. For each of the five fixed display 1 sigma levels, an average level of sigma was calculated using Levitt's reversal mean technique (Reference 12). In that technique, the stimulus level associated with P(C) = 0.72 is estimated by averaging the midpoints of either the ascending or descending series of each track. Customarily, some number of the early midpoints are discarded to allow performance to stabilize. In the present experiment, the first two ascending and two descending series of every track were eliminated prior to computing the reversal mean estimates. This corresponds to approximately the same amount of data that was deleted in the single-display condition, so that the data analyzed in both the single- and dual-display conditions represent comparable levels of training.

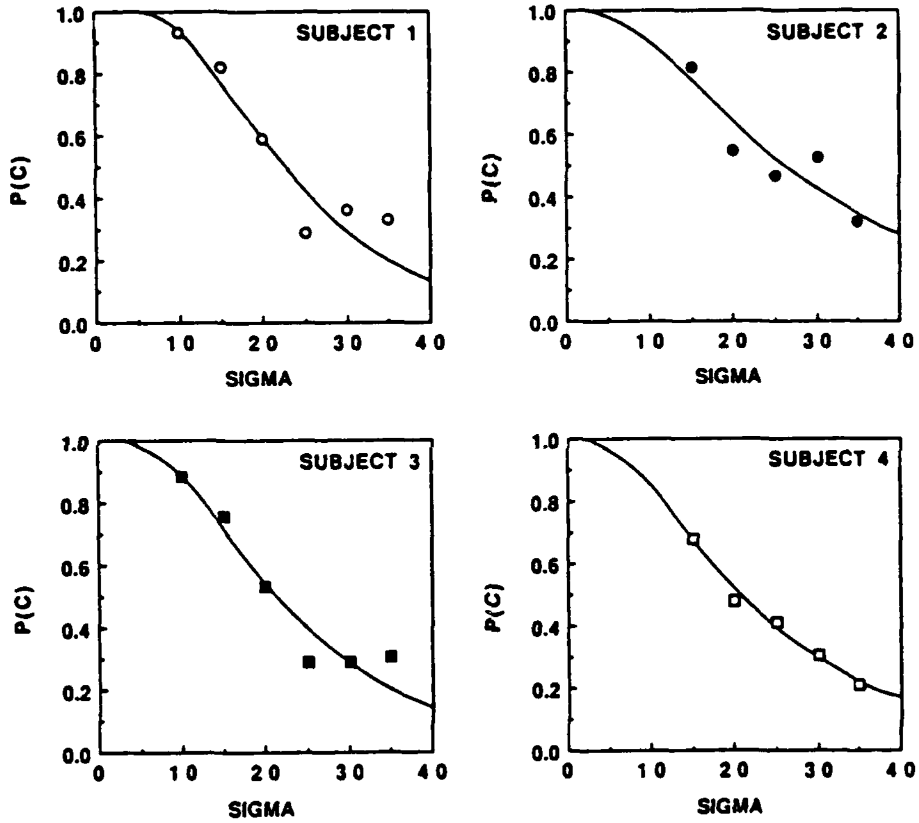


FIGURE 6. Psychometric Functions for the Four Subjects Relating Distortion Level (sigma) to Corrected-for-Chance P(C) in the Single-Display Task.

Additionally, the first few blocks of testing included relatively easy levels of the five fixed sigma levels on display 1. These first blocks were discontinued as the subject became better at the task, and were replaced by higher sigma levels. For example, the five levels for subject no. 2 ranged from 0 to 20 at the beginning of the experiment but were increased to a range of 15 to 35 to keep the stimuli tested in the appropriate meaningful psychophysical range and to not exceed the limitations of the tracking algorithm. Clearly, in an integration task, the algorithm will fail to converge when single-display performance exceeds the value that the algorithm is attempting to maintain (in this case) P(C) = 0.72 with two displays. The discontinued tracks were not included in the analysis.

When the five fixed sigma levels on display 1 were changed as described above, the tracks were continued if possible. In a small number of the tracks, the track was erroneously "reset" to an arbitrary starting value. In those cases, the ascending or descending series, which included the arbitrary starting value, was not included in the estimate of the reversal mean.

Table 1 shows the dual-display performance data for those tracks that yielded reversal mean estimates after eliminating the first two ascending and descending series.

TABLE 1. Tracking Algorithm Estimates of Display 2 Sigma Level as a Function of Display 1 Sigma Level for Each Subject in the Dual-Display Condition. SE = standard error of the mean; n = number of reversals.

Subject no.	Reversal mean estimates			
	Display 1 sigma	Display 2 sigma	SE	n
1	15	20.68	1.29	11
	20	17.00	2.06	10
	25	15.00	1.60	7
	30	11.00	1.07	10
	35	17.50	2.89	3
2	15	25.83	4.41	3
	20	27.50	0.00	1
	25	22.00	3.10	5
	30	19.38	1.20	4
	35	13.50	0.61	5
3	15	16.00	1.14	5
	20	14.38	1.03	8
	25	13.21	0.83	7
	30	8.93	0.99	7
	35	12.50	0.00	1
4	17	26.43	5.26	7
	19	16.04	1.36	12
	21	17.00	1.66	10
	23	15.91	1.23	11
	25	13.75	0.90	12

IMPLEMENTATION OF THE EVALUATION FRAMEWORK

Using the data from subject no. 4, Figure 7 is an example of the analysis method. The lower panel shows the stimulus track determined by the threshold estimation algorithm for the condition in which the distortion of the ship images in the fixed display, display 1, was sigma = 25. Based on the pattern of this subject's responses, the sigma level of display 2 was either increased or decreased across trials in an attempt to maintain P(C) = 0.72 (corrected for chance). The mean of the midpoints of the 12 descending series was calculated to be 13.75, as shown by the bold line. Therefore for this subject, when sigma = 25 on display 1, the sigma level on display 2 had to be 13.75 to obtain P(C) = 0.72 (corrected for chance).

These two sigma levels can then be "scaled" by the single-task psychometric function as shown in the upper half of Figure 7. In this case, when display 1 was sigma = 25 it yielded P(C) = 0.39 in isolation. When display 2 was sigma = 13.75 it yielded P(C) = 0.71 in isolation. When both displays were available, however, performance was presented simultaneously with a P(C) of 0.72. That is, for this condition and subject, in P(C) accuracy terms, a display "worth" 0.39 plus one "worth" 0.71 combined to be "worth" 0.72.

All of the data pairs yielding iso-performance levels can be scaled in the manner described above and plotted on the evaluation framework graph. Figure 8 shows the data in P(C) units as scaled by the P(C) psychometric functions, and all points representing P(C) = 0.72 (corrected for chance) dual-display performance. Figure 9 shows the data in d' units as scaled by the d' psychometric functions, with all points representing d' = 1.86 (which corresponds to P(C) = 0.72). In both figures, only the data based on three or more reversals (n > 3 from Table 1) are plotted. The two curves represent predictions of the two optimal integration models (statistical summation and observation integration) as described by the equations shown in the figures. The d' predictions and P(C) predictions, respectively, were converted to P(C) and d' units according to algorithm no. 2 in Reference 15.

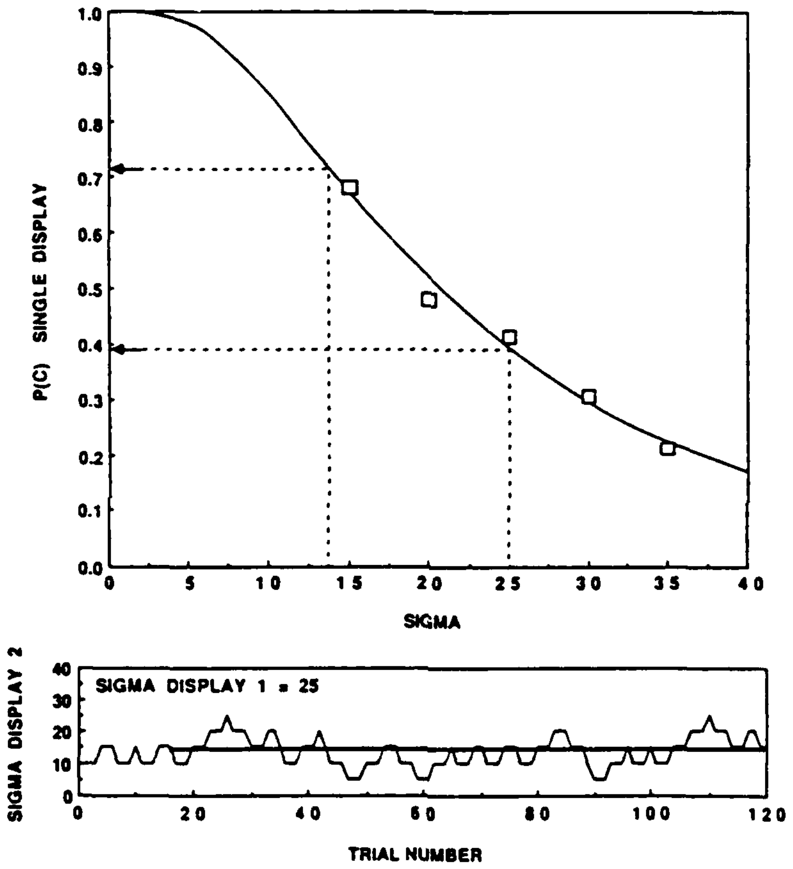


FIGURE 7. Example of the Performance Scaling Method, in Which Dual-Display Performance is Calculated in Terms of Single-Display Performance.

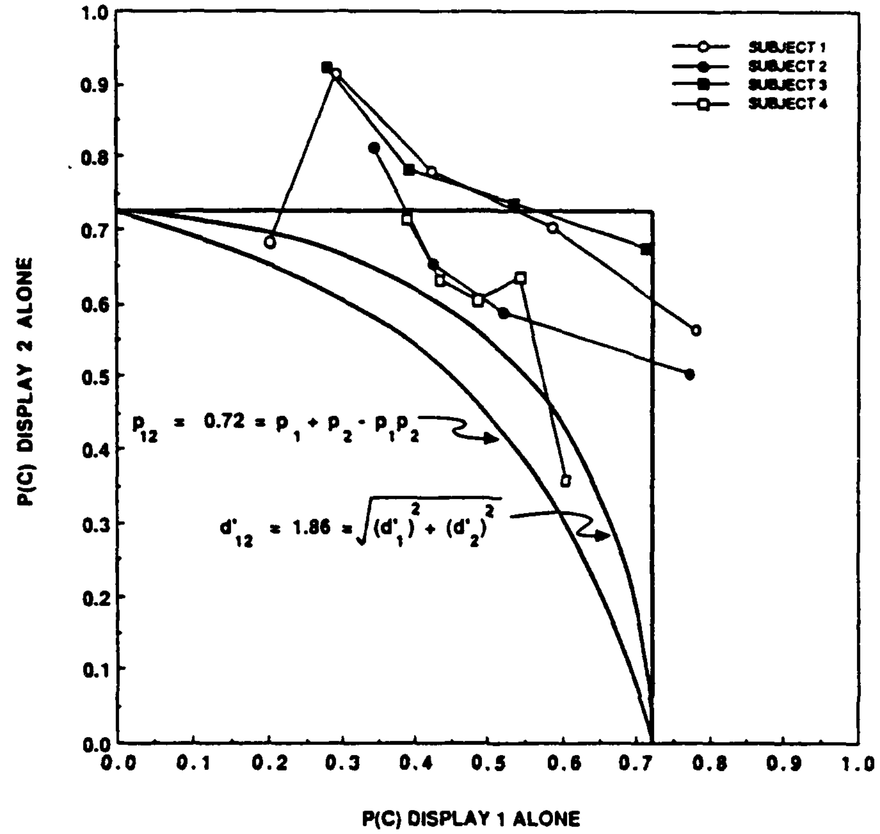


FIGURE 8. Experimental Data, in Corrected-for-Chance P(C), Overlayed on the Proposed Evaluation Framework.

DISCUSSION AND CONCLUSIONS

Ten of the eighteen data points in Figures 8 and 9 lie in the triangular "performance enhancement" region when plotted onto the evaluation framework graph. For those conditions, the subjects were able to integrate the images from the two displays and performed better than when only one of those displays was available. The conditions that led to integration appear to occur when display no. 1 was of moderate distortion (approximately P(C) = 0.50 in Figure 8, and d' = 1.25 in Figure 9).

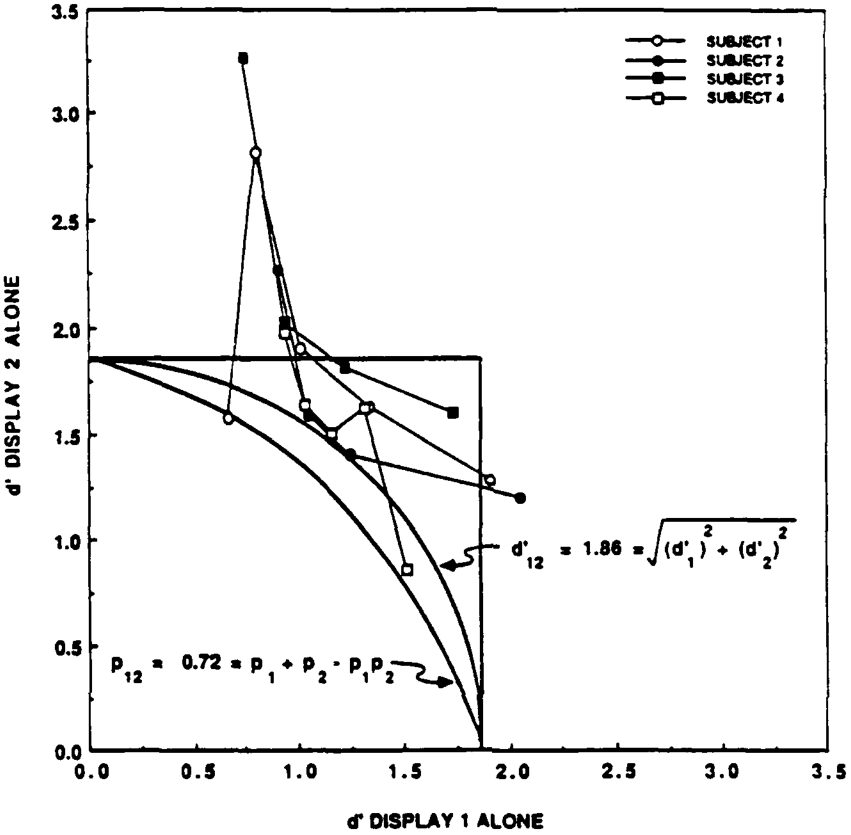


FIGURE 9. Experimental Data, in d', Overlayed on the Proposed Evaluation Framework.

When a highly distorted display (yielding about P(C) = 0.30) is presented as display no. 1, the images in display no. 2 must be of very low distortion to yield P(C) = 0.72 with both displays. In fact they must be of such low distortion that if presented in isolation, they would have yielded a performance of P(C) = 0.80 or 0.90. The subjects would have done better in those conditions if the subjects had simply ignored the highly distorted images on display no. 1 and based their responses only on the images of display no. 2. (Graphically, that would have forced the data points onto the horizontal straight lines shown in Figures 8 and 9.)

A model in which subjects always give equal weight to the information in the two displays (despite the distortion level) would explain this finding. The effect may be similar to that noted in Reference 16 where subjects weighted obviously irrelevant information equally with relevant information. The conditions which facilitate the integration of display information, and those that not only do not facilitate, but actually decrease performance, clearly warrant more investigation.

As stated earlier, the statistical summation and observation integration models can be viewed as an upper bound to normal (not Performance Super-Enhancement) information integration. In this particular experiment, the model predictions were not only an upper bound on performance in general, but in fact were appropriate predictions since the information in the dual-display condition was independent and uncorrelated. The models' failure to predict the data establishes the existence of the subjects' cognitive limitations in this particular task.

As noted previously, since these models have both been extended in the literature to include correlation between information sources, one could use model fits to determine the level of informational redundancy between two displays. If one were able to measure objectively such redundancy on a specific task, then it would be necessary to use those equations in the evaluation framework instead of those used in Figures 8 and 9, which assume no correlation. Increasing the correlation from zero to one in these models forces the model predictions towards the vertical and horizontal lines representing the boundary of the Performance Decrement and Performance Enhancement areas in the evaluation framework. This holds because data points away from the origin represent less integration, and with increasing correlations one expects to observe less integration (adding a second display does not add as much new information with larger correlations).

In summary, the evaluation framework developed herein has been demonstrated to be a useful tool to evaluate an operator's ability to integrate information from two displays. Similarly, it has been shown how one can determine the amount of information that an operator can extract from a sensor fusion, or multisensor, display. The techniques discussed here allow the evaluation of multisensor displays by comparing multisensor display performance to the predictions of existing optimal integration models and to multiple display presentations. This evaluation allows the human factors engineer to recognize in both an absolute and a relative sense whether the proposed multisensor display does what it was designed to do, i.e., integrate the sensor information and present it well.

REFERENCES

1. Green, D.M. "Detection of Multiple Component Signals in Noise," *J. Acoustical Society of America*, Vol. 30 (1958), pp. 904-11.

2. Craig, A., W. P. Colquhoun, and D. W. J. Corcoran. "Combining Evidence Presentea Simultaneously to the Eye and Ear: A Comparison of Some Predictive Models," *Perception & Psychophysics*, Vol. 19 (1976), pp. 473-84.

3. Green, D.M., and J. A. Swets. *Signal Detection Theory and Psychophysics* . New York, Krieger, 1974. (Originally published in New York, Wiley & Sons, 1966.)

4. Swets, J.A. "Mathematical Models of Attention," in *Varieties of Attention*, ed. by R. Parasuraman and D. R. Davies. New York, Academic Press, 1984. Pp. 183-242.

5. Pirenne, M.H. "Binocular and Uniocular Thresholds in Vision," *Nature*, Vol. 152 (1943), pp. 698-99.

6. Tanner, W.P., Jr. "Theory of Recognition," *J. Acoustical Society of America*, Vol. 28 (1956), pp. 882-88.

7. Kinchla, R.A., and C. E. Collyer. "Detecting a Target Letter in Briefly Presented Arrays: A Confidence Rating Analysis in Terms of a Weighted Additive Effects Model," *Perception & Psychophysics*, Vol. 16 (1974), pp. 117-22.

8. Kinchla, R.A. "The Measurement of Attention," in *Attention and Performance VIII*, ed. b, R.S. Nickerson. Hillsdale, N.J., Erlbaum, 1980. Pp. 213-37.

9. Julesz, B. *Foundations of Cyclopean Perception*. Chicago, Ill., University of Chicago Press, 1971.

10. *Jane's Fighting Ships (1982-1983)*. Jane's Publishing Company, Ltd., J. M. Moore, ed. London, England, 1982.

11. Robinson, D.E., and L. A. Jeffress. "Effect of Varying the Interaural Noise Correlation on the Detectability of Tonal Signals," *J. Acoustical Society of America*, Vol. 35 (1963), pp. 1947-52.

12. Levitt, H. "Transformed Up-Down Methods in Psychoacoustics," *J. Acoustical Society of America*, Vol. 49 (1971), pp. 467-77.

13. Foyle, D.C., and C. S. Watson. "Stimulus-Based Versus Performance-Based Measurement of Auditory Backward Recognition Masking," *Perception & Psychophysics*, Vol. 36 (1984), pp. 515-22.

14. Egan, J.P., W. A. Lindner, and D. McFadden. "Masking-Level Differences and the Form of the Psychometric Function," *J. Acoustical Society of America*, Vol. 37 (1966), p. 1181.

15. Smith, J.E.K. "Simple Algorithms for M-Alternative Forced-Choice Calculations," *Perception & Psychophysics*, Vol. 31 (1982), pp. 95-96.

16. Tversky, A. and D. Kahneman. "Judgment Under Uncertainty," *Science*, Vol. 185 (1974), 1124-30.

INITIAL DISTRIBUTION

6 Naval Air Systems Command
AIR-5004 (2)
AIR-5313 (2)
AIR-5462, J. Cronin (1)
AIR-933G (1)
5 Chief of Naval Operations
OP-098 (1)
OP-55 (1)
OP-59C (1)
OP-982 (1)
OP-987 (1)
2 Chief of Naval Research, Arlington
OCNR-10 (1)
OCNR-20 (1)
1 Space and Naval Warfare Systems Command (SPAWAR-005)
3 Naval Sea Systems Command
SEA-61R (1)
Technical Library (2)
1 Air Test and Evaluation Squadron 4, Point Mugu (Technical Library)
1 Air Test and Evaluation Squadron 5, China Lake (Technical Library)
2 Naval Academy, Annapolis (Director of Research)
4 Naval Air Development Center, Warminster
Code 60B5, LCdr T. Singer (1)
Code 6021 (1)
Code 6022 (1)
Technical Library (1)
1 Naval Air Station, Corpus Christi (Chief of Naval Air Training Staff)
3 Naval Air Test Center, Patuxent River
Code SY-70 (1)
Code SY-72 (1)
Technical Library (1)
1 Naval Avionics Center, Indianapolis (Technical Library)
1 Naval Health Research Center, San Diego
3 Naval Ocean Systems Center, San Diego
Grossman (1)
Human Factors Group (1)
Technical Library (1)
4 Naval Postgraduate School, Monterey
J. Lind (3)
Technical Library (1)
1 Naval Research Laboratory (Technical Library)
1 Naval Surface Warfare Center, Dahlgren (Technical Library)
1 Naval Surface Warfare Center, White Oak Laboratory, Silver Spring (Technical Library)
3 Naval Training Equipment Center, Orlando
Code N-08 (1)
Code N-71 (1)
Technical Library (1)
1 Naval War College, Newport (Technical Library)
3 Office of Naval Technology, Arlington (ONT-222, Dr. S. Collyer)
3 Pacific Missile Test Center, Point Mugu
Code 1226 (2)
Technical Library (1)
1 Headquarters, U. S. Army, Fort Belvoir (Library)
1 Army Training and Doctrine Command, Fort Monroe (Technical Library)
1 Army Aeromedical Research Laboratory, Fort Rucker (Technical Library)
1 Army Human Engineering Laboratory, Aberdeen Proving Ground

1 Army Materiel Systems Analysis Activity, Aberdeen Proving Ground (AMXSY-J, Library)
1 Air Force Systems Command, Andrews Air Force Base (Library)
1 Air Force Aeronautical Systems Division, Wright-Patterson Air Force Base (Technical Library)
1 Air Force Intelligence Agency, Bolling Air Force Base (AFIA/INTAW, Maj. R. Esaw)
1 Air Force Medical Research Laboratory, Wright-Patterson Air Force Base (Code HEA)
1 Air Force Munition Systems Division, Eglin Air Force Base (MSD/ENYW, Technical Library)
1 Air Force Wright Research and Development Center, Wright-Patterson Air Force Base (Code AAWD-1)
1 Defense Intelligence Agency (Technical Library)
12 Defense Technical Information Center, Alexandria
1 Under Secretary of Defense for Research and Engineering (R&AT)
10 Ames Resarch Center (NASA), Moffett Field, CA (Dr. D. C. Foyle)
1 Bureau of Reclamation, Loveland, CO (S. Wilson)
1 Calspan Corporation, Advanced Technological Center, Buffalo, NY
1 Dr. Charles Greening, Fullerton, CA
1 Hudson Institute, Incorporated, Center for Naval Analysis, Alexandria, VA (Technical Library)
1 Hughes Aircraft Company, Los Angeles, CA (Human Factors/Displays)
1 Human Performance, Incorporated, Goleta, CA
1 Institute for Defense Analyses, Alexandria, VA (Technical Library)
1 Martin-Marietta Aerospace, Orlando, FL (Technical Library)
2 McDonnell Douglas Missile Systems Company, St. Louis, MO (Engineering Psychology)
1 Northrop Corporation, Aircraft Division, Hawthorne, CA (Technical Library)
1 Rand Corporaton, Santa Monica, CA (N. Crawford)
2 The Boeing Company, Seattle, WA
  Crew Systems (1)
  MS 41-44 (1)
1 The Johns Hopkins University, Applied Physics Laboratory, Laurel, MD (Technical Library)
1 Virginia Polytechnic Institute and State University, Blacksburg, VA (Industrial Engineering Department)

ON-CENTER DISTRIBUTION

1 Code 01
1 Code 06A
1 Code 30
1 Code 31
1 Code 3103
1 Code 3104
1 Code 3105
1 Code 3107
1 Code 311
1 Code 315
1 Code 31501
1 Code 31504
1 Code 31506
1 Code 3151
10 Code 3152 (Kibbe)
1 Code 3153
1 Code 3154
1 Code 3156
1 Code 3158
1 Code 3159
1 Code 319
1 Code 32
4 Code 343 (3 plus Archives Copy)
2 Code 3686, GIDEP
1 Code 38
1 Code 39